

Conference Abstract

KIEBIDS: A Modular Framework for Scalable Information Extraction from Written Records in Natural History Collections

Christian Bölling[‡], Théo Léger[‡], Joachim Snellings[‡], Margot Belot[‡], Franziska Schuster[‡]

[‡] Museum für Naturkunde - Leibniz-Institute for Evolution and Biodiversity Science, Berlin, Germany

Corresponding author: Christian Bölling (christian.boelling@mfn.berlin)

Received: 22 Dec 2025 | Published: 23 Dec 2025

Citation: Bölling C, Léger T, Snellings J, Belot M, Schuster F (2025) KIEBIDS: A Modular Framework for Scalable Information Extraction from Written Records in Natural History Collections. Biodiversity Information Science and Standards 9: e183295. <https://doi.org/10.3897/biss.9.183295>

Abstract

Natural history collections preserve invaluable records of biodiversity across time and space. Each specimen is typically accompanied by one or more labels documenting provenance, locality, and contextual data. Mobilizing this information is crucial for research on biodiversity change, biogeography, and taxonomy. To date, much of this data remains inaccessible for computational knowledge engineering approaches because manually processing and converting the sources into structured, interoperable data formats is a labor-intensive challenge due to the volume, heterogeneity, and complexity of the documents—and curatorial resources for fulfilling these tasks are generally insufficient. Artificial Intelligence (AI)-based methodologies can significantly accelerate this process and make it economically scalable.

We present KIEBIDS^{*1}, an open-source framework for specifying and executing AI-based workflows for information extraction from specimen label images. Following a linear data-pipeline architecture, workflows comprise five sequential functional steps for information extraction:

1. image pre-processing (to prepare input images for subsequent analysis),
2. layout analysis (to identify image regions that are relevant for information extraction),

3. optical character recognition (to identify text on the syntactical level),
4. semantic parsing (to identify text that references categories of interest),
5. entity linking (to identify entities of interest mentioned in the text with authority records).

Modularity and adaptability are central design principles for the framework's architecture. Each function can be realized by one or more modules that operate independently through file-based input and output, enabling substitution or extension as new technologies emerge. This ensures flexible adaptation to various information extraction goals or new data domains.

In the current [release](#), image pre-processing is implemented using the [OpenCV](#) framework with steps for resizing, grayscale conversion, noise reduction, and binarization. Layout analysis, based on the [Segment Anything Model](#), identifies image regions that depict labels. Character recognition is implemented using two alternative modules. Besides [EasyOCR](#), [Moondream](#) is used to leverage locally-deployable vision-language model (VLM) technology. Semantic parsing is implemented using [spaCy](#) and regular expressions, as rule-based parsing has proven efficient for syntactically well-defined entities, such as dates or coordinates, given the sparse context of label texts. Entity linking, in the current release, is realized for geographical place names using the [GeoNames](#) application programming interface (API).

The input for a given workflow run consists of document images and configuration parameters. The configuration parameters encompass

1. settings for the pipeline as a whole, such as location of input and output files and execution mode,
2. the configurable settings for each functional step of the pipeline, e.g., models to be used, model parameters or the tag selection for the semantic tagging.

The workflow's output are [PAGE-XML](#) files containing image annotations, including the extracted and annotated text. Optionally, intermediate data and evaluation metrics can be assessed. Integrating seamlessly with Python codebases, [Prefect](#) is used for scheduling, monitoring, and graphical user interaction.

By combining existing open frameworks rather than developing new components, the project leverages recent advances in computer vision and natural language processing to mobilize biodiversity data. Future developments will focus on improving user experience, integrating better models for handwritten text, and expanding semantic analysis capabilities. KIEBIDS' source code^{*2} is openly available and locally deployable with moderate hardware requirements.

Keywords

collection digitization, natural language processing, OCR, transcription, collection labels, semantic parsing, named entity recognition, data pipeline

Presenting author

Christian Bölling

Presented at

Living Data 2025

Acknowledgements

This work was supported by the [AI Eco lab](#), an initiative of the German Federal Ministry for the Environment, Climate Action, Nature Conservation and Nuclear Safety.

Conflicts of interest

The authors have declared that no competing interests exist.

Endnotes

*1 "KIEBIDS" is the acronym for the German phrase "KI-basierte Extraktion Biodiversitätsrelevanter Daten aus Schriftgut" which translates to "AI-based extraction of biodiversity-relevant data from written sources".

*2 <https://github.com/MfN-Berlin/app-kie bids>